DOI:
10.1007/s00355-016-0972-1

Document Version
Peer reviewed version

Decision Sciences and the New Case for Paternalism:

Three Welfare-Related Justificatory Challenges

Article Forthcoming in *Social Choice & Welfare*

Abstract

Several authors have recently advocated a so-called new case for paternalism, according to which empirical findings from distinct decision sciences provide compelling reasons in favour of paternalistic interference. In their view, the available behavioural and neuro-psychological findings enable paternalists to address traditional anti-paternalistic objections and reliably enhance the well-being of their target agents. In this paper, I combine insights from decision-making research, moral philosophy and evidence-based policy evaluation to assess the merits of this case. In particular, I articulate and defend three complementary arguments that, I claim, challenge even the best available calls for such case. In doing so, I identify the main justificatory challenges faced by the new paternalists and explicate the implications of these challenges for the ongoing philosophical debate about the justifiability of paternalistic interference.

Keywords: Decision-Making; Welfare; Paternalism; Moral Justification; Evidence-Based Policy Evaluation.

Word Count: 9198

Introduction

In recent years, several authors have advocated a so-called new case for paternalism (henceforth, NCP), according to which empirical findings from distinct decision sciences provide compelling reasons in favour of paternalistic interference (see e.g. Hausman and Welch, 2010, and Rizzo and Whitman, 2009a, for detailed reconstructions). The idea is that the available behavioural and neuro-psychological findings enable paternalists to address traditional anti-paternalistic objections and reliably enhance the well-being of their target agents. Behavioural and neuro-psychological findings have been claimed to support paternalistic interference in a variety of domains, ranging from consumer choices to health care and reproductive decisions. These claims, in turn, prompted heated discussions regarding the justifiability of paternalism both in philosophy (see e.g. Bovens, 2009, and Carter, 2014) and in other disciplines (see e.g. Rubinstein and Arad, 2015, Sugden, 2008, and the special issue of this Journal, 2012, 38 (4), in economics; Glaeser, 2006, and Rachlinski, 2003, in psychology; and Camerer, 2006, and Farah, 2012, in neuroscience).

In this paper, I combine insights from decision-making research, moral philosophy and evidence-based policy evaluation to assess the merits of the NCP. The paper proceeds as follows. In *Section 1*, I identify and discuss three distinctive features of paternalistic interference. In *Section 2*, I reconstruct the NCP and explicate how recent behavioural and neuro-psychological findings supposedly support it. In *Sections 3-5*, I articulate and defend three complementary arguments that, I claim, challenge even the best available calls for the NCP. More specifically, the *argument from conceptual ambiguity* builds on major difficulties inherent in defining and measuring well-being to question the new paternalists' ability to show that their interventions reliably enhance agents' well-being. The *argument from limited overlap* points to the paucity of paternalistic interventions that reliably enhance agents' well-being without involving morally objectionable violations of these agents' autonomy or consent. The *argument from constrained epistemic access* aims to demonstrate that the new paternalists typically lack the information required to design and implement welfare-enhancing paternalistic interventions. These three arguments do not license all-encompassing opposition to paternalism. Still, taken together, they cast serious doubts on the new paternalists' claim that recent findings from the decision sciences provide convincing reasons for paternalistic interference.

The expression 'new paternalism' is often used to encompass a broad set of paternalistic proposals, including so-called asymmetric paternalism, which aims to enhance the well-being of 'boundedly rational' individuals while minimizing the welfare losses for 'fully rational' individuals (see e.g. Camerer et al., 2003), and libertarian paternalism, which aims to steer individuals' behaviour in welfare-enhancing directions without restricting the range of options available to such individuals (see e.g. Thaler and Sunstein, 2003). Below I adopt this entrenched use of the expression 'new paternalism' unless specified otherwise. My evaluation of the NCP aims to inform the contemporary discussion regarding the justifiability of paternalism in at least three respects of general interest to economists, philosophers and policy makers. First, it elucidates the

epistemic and evidential relevance of recent findings in economics, psychology and neuroscience for the ongoing philosophical debate about the justifiability of paternalistic interference. Second, it highlights in what respects the new paternalists improve over previous calls for paternalistic interference and explicates the main justificatory challenges faced by the NCP. And third, it brings together parallel debates that are insufficiently integrated across philosophy and specific decision sciences to develop a systematic appraisal of the NCP. In articulating this appraisal, I provide examples from a wide range of economic and policy contexts, as opposed to one single case study. I do so to make clear that my challenges apply not merely to a few selected paternalistic policies, but rather generalize across the new paternalists' policy proposals.

1. Paternalism: Distinctive Features

The notion of paternalism has been given several characterizations by philosophers (see e.g. Arneson, 1980 and 2005, Feinberg, 1971 and 1986, ch.1, Dworkin, 1972 and 1983, and Shiffrin, 2000). I am not concerned here with assessing these characterizations or with proposing a novel characterization of paternalism. For the purpose of this paper, I employ the term 'paternalistic' to indicate interventions which: (1) violate (or interfere with) the autonomy of their target agents; (2) are implemented without the explicit consent of these agents; and (3) are designed with the primary aim to enhance the well-being of those agents. This characterization singles out three features as distinctive of paternalistic - as opposed to non-paternalistic - interventions. These features, in turn, are taken to constitute individually necessary and jointly sufficient conditions for regarding an intervention as paternalistic. A few clarifications about these three conditions are in order.[1]

Concerning condition (1), paternalistic interventions greatly differ in the extent to which they violate the autonomy of their target agents (e.g. compare coercive detention with a mildly manipulative social advertising campaign). Indeed, some interventions involve such limited violations that one may question whether they are plausibly regarded as paternalistic (see e.g. Mitchell, 2005, on various forms of rational persuasion). Still, as I illustrate in *Section 2*, autonomy violations are one of the main reasons why paternalistic interventions are often deemed to be morally objectionable.[2] As to condition (2), for an intervention to

[1] Some authors put forward different characterizations of paternalism (e.g. Shiffrin, 2000, holds that not all paternalistic interferences aim to enhance the well-being of the targeted agents, and Sunstein and Thaler, 2003, do not regard violations of agents' autonomy as a necessary condition for counting interventions as paternalistic). Still, I take the tripartite characterization in the text to be sufficiently precise for the purpose of my evaluation and sufficiently general to cover many entrenched characterizations of paternalism (see e.g. Dworkin, 2010, New, 1999, and Wilson, 2011).

[2] I am not concerned with assessing how the notion of autonomy is most aptly conceptualized. For my evaluation, it suffices to note that many paternalists and anti-paternalists alike hold that individuals have an interest in deliberating and acting in light of considered judgments about their own well-being, and that this interest is plausibly understood as an interest in autonomy (see e.g. Dworkin, 1988, ch.1, and Hausman and Welch, 2010). Some works relate paternalism to interventions that violate the freedom of choice (rather than the autonomy) of their target agents (see e.g.

qualify as paternalistic it is not required that its target agents actively oppose it. Rather, a lack of explicit consent by the time the intervention is implemented is sufficient to satisfy condition (2). This construal of condition (2) covers both cases where the target agents, unbeknown to the choice architects, express their consent to some other party, and cases where these agents do not consent but would have consented had they been better informed (hypothetical consent) or ideally rational (ideal consent).[3] Finally, condition (3) relates paternalism to interventions that primarily aim to enhance their target agents' well-being, as judged by the choice architects or by those agents themselves. Below I employ the expression 'welfare-enhancing' as a place-holder for different conceptions of well-being, without taking a position as to how exactly well-being should be defined and measured. In particular, I claim that an intervention is 'welfare-enhancing' if it improves the well-being of its target agents with respect to an otherwise identical situation where such intervention is not implemented. This use of the expression 'welfare-enhancing' encompasses both situations where some other non-paternalistic intervention is implemented and situations where no other intervention is implemented.

On the outlined construal of conditions (1)-(3), whether an intervention can qualify as paternalistic depends on whether such intervention is designed with the primary aim of enhancing the well-being of its target agents, not whether it succeeds in achieving this aim. That is to say, interventions that happen to make their target agents worse off can count as paternalistic if they were designed with the primary aim of benefiting those agents (see e.g. Bullock, 2015, on similar characterizations of paternalism). Moreover, interventions designed with additional aims besides that of enhancing agents' well-being may still qualify as paternalistic (see e.g. Dworkin, 2005, on moral paternalism). Establishing whether enhancing agents' well-being is the primary aim with which an intervention is designed is not always straightforward (e.g. policy interventions are often designed with multiple aims). However, this complication does not affect my evaluation, since most new paternalists emphasize welfare enhancement as the primary aim of the interventions they advocate (see e.g. Sunstein and Thaler, 2003, 1159; see also Diener and Seligman, 2004, 1-2, for the claim that well-being "ought to be the ultimate goal around which economic, health, and social policies are built").

2. The New Case for Paternalism

The NCP draws on a wide array of recent behavioural and neuro-psychological findings. According to the new paternalists, these findings enable choice

---

Carter, 1999, ch.8, and 2014). I mention only autonomy in the text for expositional convenience. I take my critique of the NCP to hold *mutatis mutandis* for characterizations of paternalism that relate it to interventions that violate agents' freedom of choice (rather than autonomy).

[3] Several questions arise concerning the notion of consent (e.g. what circumstances license inferring that an agent consents to a particular interference? Under what conditions does consent count as fully informed or ideally rational?). I do not expand on these issues since the cogency of my evaluation does not rest on what position one holds about them (for a recent discussion, see e.g. Groll, 2012, and Husak, 2010).

architects to improve over former calls for paternalistic interference with regard to each of the three distinctive features outlined in *Section 1*. The idea is to build on the available evidence to implement paternalistic interventions which respectively (1) involve morally acceptable violations of (or interferences with) the autonomy of their target agents, (2) harmonize with these agents' hypothetical or ideal consent, and (3) reliably enhance the well-being of those agents. Let us consider these three alleged improvements in turn.[4]

(1) Autonomy-related concerns figure prominently in the writings of anti-paternalists (see e.g. Kant, 1797 [1996], MM 6:453). One common theme in this literature is that even if individuals fail to make welfare-enhancing choices, the welfare losses they incur do not justify third parties' interferences, for "autonomy is even more important than personal well-being" (Feinberg, 1986, 59). To be sure, most anti-paternalists concede that some paternalistic interferences (e.g. primary schooling) involve morally acceptable violations of agents' autonomy. In particular, few authors take any minor violation of autonomy to *ipso facto* make the associated paternalistic interferences unjustified. Still, autonomy violations are one of the main reasons why paternalistic interferences are often deemed to be morally objectionable. In the words of Velleman, "the reasons for deferring to a person's judgment […] go beyond his reliability as a judge. Respect for a person's autonomy may require that we defer to his considered judgment […] even when we have reason to regard that judgment as mistaken" (1999, 608; see Darwall, 2006, 280, for a similar remark).[5]

The proponents of the NCP advocate a range of paternalistic interventions that putatively enhance agents' well-being without involving morally objectionable violations of these agents' autonomy. For example, Sunstein and Thaler (2003) support mandatory cooling-off periods that aim to benefit agents by inducing them to critically reconsider their own decisions. For his part, Cohen (2013) argues that so-called active choice, which requires agents to make a decision before a specified deadline, improves agents' well-being in several circumstances (e.g. urgent decisions about medical treatment). In recent years, various authors have invoked autonomy-related concerns not so much against, but rather in favour of paternalistic interference. The idea is that autonomy involves not merely having one's preferences protected from undesirable influences, but also being able to deliberate and act in light of considered judgments concerning one's well-being (see e.g. De Marneffe, 2006). On this basis, some proponents of the NCP support paternalistic interventions that - while involving temporary violations of agents' autonomy - counteract the

[4] The new paternalists typically allege that the interventions they advocate are superior to traditional paternalistic interventions in each of these three respects, individually considered, but rarely examine how such interventions fare in all those respects, collectively considered. I explore this issue's implications for the NCP in *Section 4*.

[5] Anti-paternalists may draw on both deontological and consequentialist considerations to support these autonomy-related concerns. By way of illustration, suppose facing some agents engaged in self-regarding conduct that has no direct and significant effects on the well-being of others. A consequentialist may argue that since many agents value the opportunity to make autonomous decisions, and since giving agents opportunities they value often enhances their well-being, paternalistic interventions that frustrate this opportunity rarely turn out to be welfare-enhancing (see e.g. Sugden, 2004).

influence of factors that purportedly impede autonomous decision-making. For instance, Thaler and Sunstein (2008, ch.2) advocate restricting the short-term range of options of specific classes of agents (e.g. addicts) on the alleged ground that doing so would safeguard or even promote these agents' long-term autonomy.[6]

(2) The mere fact that an agent does not explicitly consent to a particular intervention does not imply that such intervention violates her hypothetical or ideal consent. The new paternalists frequently advocate paternalistic interventions that, while operating without individuals' explicit consent, putatively harmonize with their hypothetical or ideal consent (see e.g. Sunstein, 2013a). The declared aim of these interventions is to help individuals to achieve their own considered goals without steering their behaviour towards predetermined outcomes (see e.g. Bar-Gill and Sunstein, 2015). Such paternalistic interventions are claimed to address or circumvent anti-paternalistic concerns associated with violations of agents' consent. To give one example, paternalists are often criticized for failing to respect the preferences of their target agents. Most of these criticisms implicitly presuppose that the involved agents possess well-defined preferences before facing specific decision problems. In many cases, however, people construct their preferences only when confronting such problems (see e.g. Guala, 2005). According to some authors (e.g. Thaler and Sunstein, 2003, 1164), in these cases it is pointless to criticize paternalists for failing to respect agents' preferences. For those agents lack well-defined preferences in the first place.[7]

(3) As to the enhancement of agents' well-being, the following reasoning is often put forward by the new paternalists (see e.g. Bhargava and Loewenstein, 2015, and Loewenstein and Haisley, 2008). Paternalistic interference is commonly opposed on the alleged ground that individuals are better placed than third parties to determine what choices enhance their own well-being (see e.g. Mill, 1859 [1956], ch.3-4). Even so, individuals frequently fail to make welfare-enhancing decisions. Moreover, several factors besides a lack of information regarding the available options can lead individuals to make choices that worsen (rather than enhance) their own well-being (see e.g. Tversky and Kahneman, 1974, on cognitive biases, and Elster, 1984, part II, on self-control problems). Until recently, choice architects could influence only a narrow subset of these factors and had limited control over them. Fortunately - the reasoning goes - recent findings from the decision sciences enable choice architects to intervene on a wider range of factors and exert a more pervasive

[6] Not all conceptions of autonomy are equally hospitable to these considerations. For instance, some Kantians would presumably object that autonomous agency cannot be subjected to the instrumental considerations seemingly involved in the aforementioned intertemporal trade-offs. I expand in *Section 4* on the justificatory challenges that violations of autonomy pose to the NCP.

[7] This obviously does not preclude one from opposing such interventions on other grounds. For instance, one may act on preferences that are not stable under reflection, be aware that her preferences are unstable, and yet attribute a high importance to the opportunity to satisfy her unstable preferences (see e.g. Sugden, 2006 and 2007). Furthermore, many individuals have strong preferences against having their preference-formation mechanisms influenced by third parties' interference, and the new paternalists' interventions often frustrate such preferences (see e.g. Sugden, 2013).

control over them. This, in turn, provides choice architects with the means to design and implement paternalistic interventions that reliably enhance well-being both across agents and across tokens of interventions of the same type.

Two kinds of contributions appear to be particularly significant in this context. The first relates to the new paternalists' attempts to improve individuals' well-being by exploiting specific biases and behavioural regularities. For instance, so-called save more tomorrow plans can noticeably increase employees' savings for retirement by changing their default options in retirement saving decisions. These interventions exploit individuals' status quo bias and do not restrict the set of options available to them (see e.g. Thaler and Sunstein, 2008). The second kind of contributions concern the use of neurochemicals and hormones to alter agents' behaviour. The goal is to identify how specific neuro-physiological perturbations affect decisions in particular experimental settings and use this information to influence individuals' decisions in real-life situations. For example, various studies (e.g. Baumgartner et al., 2008, and Kosfeld et al., 2005) document that modulating individuals' oxytocin levels can significantly increase their propensity to trust other players in experimental settings where cooperative behaviour is deemed to be welfare-enhancing. Whether the interventions based on these studies qualify as paternalistic would depend on whether such interventions violate the autonomy and the consent of the targeted agents (see *Section 1*). Still, those studies provide a nice illustration of how recent neuro-psychological findings may help paternalists to foster welfare-enhancing modifications in agents' behaviour. As Farah puts it, "[in] education, business, politics, law [...] any endeavor that depends on being able to […] improve human behavior is, in principle, a potential application area for neuroscience" (2012, 57).

3. Argument from Conceptual Ambiguity

As outlined in the previous section, the NCP is premised on the assumption that the new paternalists' interventions reliably enhance the well-being of their target agents. In this section, I articulate and defend an *argument from conceptual ambiguity* that questions the new paternalists' ability to substantiate this assumption. My argument goes as follows. In the behavioural and neuro-psychological literature, different methods have been developed for measuring well-being (see e.g. Hausman, 2012, ch.7-8, and Rubinstein and Salant, 2012). Moreover, dissimilar conceptions of well-being have been proposed (see e.g. Griffin, 1986, and Parfit, 1984, 493-502, for an entrenched tripartition between mental state conceptions, preference satisfaction conceptions and objective list conceptions). The new paternalists advocate dissimilar measures and conceptions of well-being (see e.g. Le Grand and New, 2015, and White, 2013). This holds not just for different authors, but also for different works by the same authors (see e.g. Qizilbash, 2012, on distinct variants of preference satisfaction conceptions advocated by Sunstein and Thaler) and at times even for one and the same work (see e.g. Sunstein, 2013b, 1862, who uses 'welfare' to indicate both "whatever choosers think would make their lives go well" and "whatever the paternalist thinks would make choosers' lives go well"). These divergences make it hard to assess the welfare implications of paternalistic interference. In particular, they cast doubt on the new paternalists' claim that

their interventions reliably enhance agents' well-being. For in many cases, whether or not paternalistic interference can be plausibly taken to enhance agents' well-being crucially depends on what conception of well-being one endorses and on what methods one employs to measure well-being.

To illustrate this, let us consider the entrenched tripartition between mental state conceptions, preference satisfaction conceptions and objective list conceptions of well-being. Mental state conceptions hold that well-being consists in the presence of specific kinds of mental states (see e.g. Crisp, 2006, and Feldman, 1997, on hedonistic conceptions). Preference satisfaction conceptions, instead, hold that an agent is well off to the extent that her actual, informed or ideal preferences are satisfied (see e.g. Hausman and McPherson, 2009, and Sumner, 1995). An agent's preferences count as satisfied when the state of affairs with which these preferences are concerned obtains. The actualization of this state of affairs, in turn, does not have to involve any psychological feeling or experience of satisfaction on the part of the agent, and may even obtain without the agent being aware of such actualization (see e.g. Hausman, 2011). Still differently, objective list conceptions hold that certain goods or experiences contribute to an agent's well-being regardless of whether they bring about particular mental states or satisfy the agent's preferences (see e.g. Nussbaum and Sen, 1993). To be sure, various objective list conceptions allow that experiencing specific mental states and satisfying particular preferences may contribute to well-being. However, on objective list conceptions, an agent's well-being is not determined solely by the agent's own mental states and preferences (see e.g. Hausman, 2010).

These three sets of conceptions occasionally yield consistent verdicts as to whether specific types of paternalistic interference are welfare-enhancing (e.g. think of primary schooling). Nonetheless, the new paternalists' reliance on dissimilar conceptions of well-being poses a severe justificatory challenge to the NCP. For in many cases, different conceptions license conflicting evaluations of the welfare implications of paternalistic interference (see e.g. Griffin, 1986, on conflicts between mental state conceptions and objective list conceptions; Arneson, 1999, on conflicts between preference satisfaction conceptions and objective list conceptions; and XXX on conflicts between mental state conceptions and preference satisfaction conceptions). Indeed, such conflicts can occur even with distinct variants of *the same* conception (see e.g. Zamir, 1998, on how different sets of paternalistic interventions can be plausibly taken to make agents better off depending on whether one endorses an actual - as opposed to ideal - preference satisfaction conception of well-being). In this respect, it is telling that distinct authors' positions regarding the justifiability of paternalism frequently vary with what conceptions of well-being they endorse (e.g. think of many economists' anti-paternalism and their reliance on actual or informed preference satisfaction conceptions of well-being).

A proponent of the NCP may object that the new paternalists' reliance on dissimilar conceptions of well-being constitutes an unproblematic or even welcome indication of pluralism on their part (see e.g. Thaler and Sunstein, 2003). After all - the thought would be - there is widespread disagreement about the relative merits of distinct conceptions, and it would be unnecessarily

demanding to require the new paternalists to agree on a single conception of well-being. This objection provides little support to the NCP. For in many cases where distinct conceptions of well-being yield consistent *verdicts* concerning the welfare implications of paternalistic interference, these conceptions ground such verdicts on rather different *justificatory principles*. These differences, in turn, can significantly constrain the robustness of the new paternalists' agreement regarding the welfare implications of interventions that target diverse classes of agents and choice situations. Indeed, such differences may constrain not just the robustness, but also the informativeness of the new paternalists' agreement regarding the welfare implications of their interventions. To see this, consider the ongoing debate concerning the informativeness of distinct indicators of well-being (see e.g. XXX). Several authors take observed correlations between putative indicators of well-being (e.g. hedonic reports, neuro-biological variables) to show that these indicators provide accurate and reliable measures of well-being. Those correlations may well suggest that the indicators at hand target some common variable, yet do not demonstrate that such indicators provide accurate and reliable measures of well-being (see e.g. Bernheim, 2009).

A proponent of the NCP may further object that the previous remarks do not support *selective* anti-paternalism, since they apply to most interventions aimed at improving agents' well-being *irrespective* of whether these interventions qualify as paternalistic (see e.g. Sunstein, 2015). The idea would be that the divergences between distinct conceptions of well-being make it difficult to reach agreement on the welfare implications of several interventions, independently of whether these interventions are paternalistic. This objection invites the following two-fold rejoinder. First, the mere fact that paternalistic and non-paternalistic interventions face some common justificatory concerns does not exempt paternalists from the need to address these concerns (see e.g. *Section 5* on various cases where agents' self-regulatory efforts have better welfare implications than both paternalistic and non-paternalistic interventions). And second, several justificatory concerns support selective anti-paternalism, since they predominantly challenge paternalistic - as opposed to non-paternalistic - attempts to enhance agents' well-being. I shall expand in *Sections 4* and *5* on these justificatory concerns. For the purpose of this section, I outline some difficulties inherent in establishing whether paternalistic interference reliably yields welfare benefits to its target agents.

In recent years, the new paternalists have offered various criteria for establishing whether paternalistic interference yields welfare benefits to its target agents. For instance, as noted by Thaler and Sunstein (2008, ch.4), paternalistic interference is likely to be welfare-enhancing in situations where individuals obtain delayed and limited feedbacks concerning their choices' welfare implications. Unfortunately, these criteria are set at an exceedingly high level of abstraction to enable the new paternalists to show that the interventions they advocate reliably yield welfare benefits to their target agents. To illustrate this, consider the new paternalists' claim that one can establish whether their interventions are welfare-enhancing by examining *aggregate* data (see e.g. Bar-Gill and Sunstein, 2015) and the *hypothetical* decisions that the majority of agents would presumably make if explicit choices were required (see e.g. Sunstein and Thaler, 2003). These data rarely provide precise indications

concerning the welfare implications of the new paternalists' interventions across individuals and choice situations. For instance, consider the so-called endowment effect, i.e. individuals' tendency to value specific goods more if they are given initial ownership of such goods. As acknowledged by Jolls and Sunstein (2006, 220), several factors may lead to differences between individuals' willingness to accept and willingness to pay, with controversial axiological assumptions being required to establish whether these differences constitute errors in need of correction and, if so, how to correct such differences.

To give another example, take the new paternalists' declared aim to steer agents' behaviour towards the choices these agents would make "if they had *complete* information, *unlimited* cognitive abilities, and *no lack* of self-control" (Sunstein and Thaler, 2003, 1162, italics added). Achieving this aim would require the new paternalists to identify precisely what choices the targeted agents would make under these ideal conditions. This identification exercise, in turn, faces three major difficulties. First, it is hard to establish what exactly agents' complete information, unlimited cognitive abilities, and perfect willpower amount to unless one makes substantive assumptions about well-being (e.g. what information is deemed to be relevant in a given decision context can vary remarkably depending on what conception of well-being one endorses). Second, it remains obscure on what evidential and epistemic basis the new paternalists are to identify what the targeted agents would choose under ideal conditions. To be sure, various methods have been developed to reconstruct what preferences agents would exhibit if they had complete information and were free of reasoning imperfections (see e.g. Bernheim and Rangel, 2007, and Salant and Rubinstein, 2008). Yet, these methods yield informative welfare rankings only in choice situations where agents can be plausibly assumed to possess latent preferences that satisfy standard consistency principles (e.g. context independence), and it is dubious that agents generally possess such preferences (see e.g. Infante et al., 2016). And third, there is no guarantee that the latent preferences thus reconstructed provide reliable insights concerning agents' well-being. As Hausman puts it, "it is one thing to determine what people's preferences would be if they had [complete information] and were free of [reasoning imperfections], and it is a different thing to determine what is good for people" (2016, 30).

These difficulties do not prevent the new paternalists from developing *increasingly* precise criteria for evaluating the welfare implications of their interventions. Even so, the new paternalists' attempts to specify criteria that are *sufficiently* precise to provide choice architects with informative practical guidance are vulnerable to severe objections. By way of illustration, consider Sunstein and Thaler's (2003) presumption in favour of policies that minimize the number of opt-outs. The mere fact that some policy minimizes the number of opt-outs by no means implies that this policy enhances the well-being of its target agents. For the number of opt-outs associated with that policy may vary depending on several factors that are unrelated to the welfare implications of such policy. For example, the mere fact that the rate at which adolescents start smoking nicotine is higher than the rate at which smokers of the same age quit falls short of implying that smoking enhances adolescents' well-being. For this behavioural pattern more plausibly results from nicotine's addictive properties

than from the alleged fact that smoking enhances adolescents' well-being (see e.g. Sugden, 2008). Moreover, paternalistic interventions often contribute to determining what choices are subsequently made by their target agents, leading these agents to make rather different choices than the ones they would have made in the absence of such interventions (see e.g. Archard, 1993). In these situations, the observed number of opt-outs is more aptly regarded as a byproduct of paternalistic interference than as a reliable source of evidence regarding the welfare implications of such interference.

To recapitulate, the empirical findings collected in distinct decision sciences document individuals' widespread failures to make welfare-enhancing decisions, and enable the new paternalists to effect significant changes in individuals' behaviour. Nonetheless, showing that particular paternalistic interventions are welfare-enhancing typically requires one to discriminate between different conceptions of well-being, or at least provide precise and plausible criteria for evaluating the welfare implications of such interventions. In this critical respect, even the best available calls for the NCP remain grounded in exceedingly vague conceptualizations of welfare to provide choice architects with informative practical guidance.

4. Argument from Limited Overlap

According to the NCP, recent behavioural and neuro-psychological findings enable choice architects to implement paternalistic interventions which respectively (1) involve morally acceptable violations of (or interferences with) the autonomy of their target agents, (2) harmonize with these agents' hypothetical or ideal consent, and (3) reliably enhance the well-being of those agents (see *Section 2*). My *argument from limited overlap* questions the significance of these putative improvements for the merits of the NCP. The argument proceeds as follows.

Suppose, for the sake of argument, that the available behavioural and neuro-psychological findings help the new paternalists to implement paternalistic interventions that respectively satisfy conditions (1), (2) and (3), *individually* considered. This by no means implies that these paternalistic interventions satisfy *all* those three conditions, *collectively* considered. For there is no general reason to assume that the sets of paternalistic interventions that respectively satisfy conditions (1), (2) and (3) overlap to a significant extent. Now, showing that some paternalistic intervention is justified would typically require the new paternalists to demonstrate that this intervention satisfies conditions (1)-(3), collectively considered. After all, it would be of limited import to demonstrate that a paternalistic intervention enhances agents' well-being, if this intervention succeeds in doing so only by means of morally unacceptable violations of these agents' autonomy or consent. Conversely, it would hardly help the new paternalists to show that a paternalistic intervention involves morally acceptable violations of agents' autonomy and consent, if this intervention fails to reliably improve the well-being of these agents. Unfortunately, only a few paternalistic interventions are shown to satisfy conditions (1)-(3), collectively considered. This, in turn, casts doubt on the merits of the NCP. Let me explicate this point.

Most of the policy interventions advocated by the new paternalists face the following dilemma. On the one hand, several paternalistic interventions involve morally acceptable violations of their target agents' autonomy (condition 1) and consent (condition 2), but do not reliably enhance these agents' well-being (condition 3). On the other hand, other paternalistic interventions reliably enhance their target agents' well-being (condition 3), but succeed in doing so only because they involve morally objectionable violations of these agents' autonomy (condition 1) and/or consent (condition 2). This dilemma affects not just a few paternalistic interventions, but also paradigmatic types of paternalistic interference championed by many new paternalists. By way of illustration, let us consider the interventions advocated by prominent libertarian paternalists. Libertarian paternalists declaredly aim to alter their target agents' behaviour so as to "make [these agents] better off, as judged by themselves" (Thaler and Sunstein, 2008, 5). The thought is that while traditional paternalists influence agents by means such as manipulation and deception, libertarian paternalists help individuals to make welfare-enhancing choices without restricting the range of options available to them (see e.g. Sunstein and Thaler, 2003, on interventions that allow their target agents to opt out of automatic enrolments and discard the proposed default options). At first glance, the aim to make agents better off as judged by themselves without restricting the range of options available to them may seem an attractive policy ideal. Even so, the libertarian paternalists' attempts to implement interventions that respect this ideal raise at least two major concerns.

First, the mere fact that some paternalistic intervention does not restrict the range of options available to its target agents falls short of indicating that such intervention involves no morally objectionable violation of autonomy or consent (e.g. think of subliminal advertising and other forms of psychological manipulation). And second, choice architects' ability to influence agents' behaviour tends to significantly decrease when these agents are previously informed of the implementation of paternalistic interference and the cognitive mechanisms it exploits (see e.g. Bovens, 2009, and White, 2013, ch.4-5, for illustrations). Whenever this is the case, a tension arises between the choice architects' aim to enhance agents' well-being and their purported moral obligation to inform such agents of the implementation of paternalistic interference and the cognitive mechanisms it exploits. In such situations, libertarian paternalists rarely inform the targeted agents of what cognitive mechanisms are used to influence their behaviour and how exactly those mechanisms supposedly exert such influence (see e.g. Le Grand and New, 2015, ch.6-7). To see this, consider again the 'save more tomorrow' plans mentioned in *Section 2*. These interventions do not engage their target agents in a process of rational persuasion, but influence their choices by surreptitiously exploiting decision biases (e.g. status quo bias) to which they are likely vulnerable. In this respect, it is telling that leading libertarian paternalists contend that in many cases a choice architect should "make the choices that she thinks would make the [agents] best off" (Sunstein and Thaler, 2003, 1164) and "it may be desirable to impose [delegation] to protect naıve individuals who are unaware of their imperfect rationality" (Bar-Gill and Sunstein, 2015, 10).

A proponent of the NCP may reply that choice architects often *inform* the targeted agents of the implementation of paternalistic interference (Loewenstein

et al., 2015), and that providing this information suffices to satisfy basic *transparency* constraints (see e.g. Thaler and Sunstein, 2008, 244, on the so-called publicity principle, which bans governments from selecting policies that they would not be able or willing to defend publicly to their own citizens). However, the new paternalists have hitherto failed to precisely demarcate the set of interferences compatible with their transparency constraints (see e.g. Wilkinson, 2013, on Thaler and Sunstein's publicity principle). Moreover, informing the targeted agents of the implementation of paternalistic interference does not *per se* address the justificatory concerns related to the limited transparency of the new paternalists' interventions. To give one example, libertarian paternalists likely alter their target agents' behaviour in ways that elude these agents' awareness when they refrain from disclosing to those agents what cognitive mechanisms are employed to influence their behaviour and how these mechanisms supposedly exert such influence (see e.g. Felsen et al., 2013). Many agents, in turn, regard this kind of interference as more objectionable than paternalistic interventions that target conscious processes (see e.g. Rubinstein and Arad, 2015). In this perspective, several instances of libertarian paternalism resemble traditional paternalistic interventions in their tendency to substitute the choice architects' evaluations for the target agents' judgments about their own well-being.[8]

To be sure, many paternalists and anti-paternalists alike agree that if an intervention makes its target agents better off, then this fact should be regarded as a *prima facie* (albeit defeasible) reason in favour of this intervention. Moreover, few authors regard the mere fact that a paternalistic intervention involves some minor violation of autonomy or consent as a sufficient reason to oppose such intervention (see *Section 2*). Even so, paternalists and anti-paternalists respectively advocate rather dissimilar positions as to under what circumstances the welfare benefits yielded by paternalistic interference may justify such interference (e.g. are autonomy and consent just one among many factors pertaining to the evaluation of paternalistic interventions, or should they be regarded as a general constraint on the set of admissible interventions?).[9] Furthermore, the new paternalists have not specified precise and plausible criteria for assessing when exactly these welfare benefits can be taken to override the moral concerns associated with violations of autonomy and consent. This lack of specificity is problematic, since violations of autonomy and consent pose remarkable justificatory challenges to the new paternalists. Let me expand on this issue.

[8] Similar concerns arise in relation to the possibility that paternalistic interference may lead to agents' infantilization. The idea is that paternalistic interventions neither help nor incentivize their target agents to develop effective decision-making skills and make welfare-enhancing decisions for themselves (see e.g. Bovens, 2009).

[9] A new paternalist might object that if some paternalistic intervention is welfare-enhancing, then such intervention is *ipso facto* justified. However, this objection presupposes that the welfare implications of paternalistic interference include all the factors pertaining to the justifiability of such interference, and the new paternalists have not offered convincing support to this welfarist presupposition (see e.g. Kagan, 1992, and Sobel, 1998, for a discussion of the role autonomy considerations can be taken to play in the definition and measurement of well-being).

Most new paternalists acknowledge that violations of autonomy and consent may raise significant concerns about the justifiability of paternalism. Indeed, as noted in *Section 2*, several authors emphasize these concerns in highlighting the alleged superiority of the NCP over former calls in favour of paternalism. In this context, showing that choice architects should implement a given paternalistic intervention would require one to show that such intervention does not involve morally objectionable violations of its target agents' autonomy or consent and has better expected welfare implications than non-paternalistic alternatives, i.e. likely enhances its target agents' well-being with respect to situations where some alternative non-paternalistic intervention is implemented and situations where no other intervention is implemented. Regrettably, the new paternalists rarely attempt to meet this justificatory challenge. Moreover, the epistemic and evidential concerns I explicate in the next section make it highly doubtful that the new paternalists' interventions have better expected welfare implications than non-paternalistic alternatives.

5. Argument from Constrained Epistemic Access

In this section, I articulate and defend an *argument from constrained epistemic access* which aims to demonstrate that the new paternalists typically lack the information required to design and implement welfare-enhancing paternalistic interventions. If correct, this argument poses a major justificatory challenge to the NCP, since the NCP crucially rests on the assumption that the new paternalists' interventions reliably enhance the well-being of their target agents (see *Section 2*). The argument points to some major epistemic and evidential concerns that make it difficult for the new paternalists to accurately: (1) quantify the impact specific decision biases and limitations have on agents' behaviour; (2) calibrate their interventions for the interactions between these biases and limitations; (3) anticipate how responsive such biases and limitations will be to their interventions; and (4) estimate the effects of agents' self-regulative efforts on their own behaviour. These concerns do not exclude that the new paternalists might occasionally obtain the information required to design and implement welfare-enhancing paternalistic interventions. Still, taken together, they give powerful reasons to think that the new paternalists rarely possess such information and will not obtain it in the near future. Below I examine these epistemic and evidential concerns in turn and support my critique with a series of examples from economic and policy contexts. Some of those concerns affect both paternalistic and non-paternalistic attempts to enhance agents' well-being. Others, instead, prevalently challenge paternalistic interventions and single out paternalistic - as opposed to non-paternalistic - interventions as especially problematic.[10]

[10] Other authors (e.g. Glaeser, 2006, and Rizzo and Whitman, 2009b) put forward epistemic and evidential criticisms of the NCP. My remarks agree with these informative criticisms in spirit, but are grounded in a different conceptualization of paternalism and do not imply that the information required to implement welfare-enhancing interventions is "in principle" unavailable to paternalists (Rizzo and Whitman, 2009b, 159).

(1) As noted in *Section 1*, individuals are subject to a variety of decision biases and limitations. The new paternalists could in principle design welfare-enhancing interventions without having accurate knowledge of the impact that these biases and limitations have on agents' behaviour (e.g. think of cases where different biases offset each other). Even so, implementing welfare-enhancing paternalistic interventions usually requires one to identify not just *which* biases and limitations affect her target agents, but also *what impact* such biases and limitations have on those agents' behaviour. For the impact of specific biases and limitations varies significantly across agents (see e.g. Barber and Odean, 2001, on overconfidence), periods (see e.g. Baumeister, 2002, on self-control problems), and choice situations (see e.g. Samuelson and Zeckhauser, 1988, on the status-quo bias). Due to these variations, showing that a paternalistic intervention improves agents' well-being in a particular experimental setting by no means guarantees that the same holds across agents and situations. Conversely, demonstrating that some paternalistic intervention is welfare-enhancing in most situations of a given type (e.g. retirement saving decisions) does not license the claim that such intervention is welfare-enhancing in all (or even most) token situations of such type. For the welfare implications of paternalistic interference may vary dramatically due to minor alterations in the distribution of specific biases in the target population. The following example, which concerns how optimal sin taxes may vary depending on the distribution of self-control problems, nicely illustrates this point.

Several authors advocate imposing so-called sin taxes that counteract individuals' vulnerability to present-bias and lack of willpower. O'Donoghue and Rabin (2006) propose a model where individuals choose between a composite good and a 'sin good' that is enjoyable to consume, but yields health costs or other negative consequences in the future (e.g. think of cigarettes and fatty foods). O'Donoghue and Rabin investigate how the optimal level of sin tax varies depending on the values of various parameters, including the elasticity of demand for the targeted goods, the marginal health costs associated with these goods' consumption, and the distribution of self-control problems in the targeted population. Let us focus on the last parameter. As illustrated by O'Donoghue and Rabin's numerical examples (2006, 1836-9), minor variations in the distribution of self-control problems can have a dramatic effect on the optimal level of sin tax (e.g. marginal increases in agents' present-bias can lead to an increase of the optimal tax level from 5% to 63% approximately). This, in turn, is problematic because the new paternalists often lack the means to accurately estimate the distribution of self-control problems in the population segments they target. This problem is exacerbated by the fact that the distribution of self-control problems has been shown to vary both across individuals and across different sin goods (see e.g. Rizzo and Whitman, 2009b).

(2) The empirical findings collected in distinct decision sciences document that individuals' behaviour is often influenced by *a number of* biases and limitations *simultaneously*. The new paternalists can occasionally exploit these influences to steer agents' behaviour in welfare-enhancing directions. For instance, as shown by Jolls and Sunstein (2006), one may use the availability heuristic - which inclines one to judge events as more probable when they can be called to mind more easily - to make negative outcomes more salient and thereby counteract agents' optimism bias. At the same time, individuals' vulnerability

to multiple biases and limitations can severely complicate the task of estimating the welfare implications of paternalistic interference. For biases and limitations interact in dissimilar ways (e.g. by reinforcing and offsetting each other) across agents and choice situations. Hence, paternalistic interventions that correct only some of their target agents' biases can significantly worsen (rather than improve) these agents' well-being. By way of illustration, individuals' tendency to overestimate their future consumption can partly offset their propensity to under-save due to hyperbolic discounting, and paternalistic interventions that alleviate only agents' overestimation bias tend to aggravate their under-saving bias (see e.g. Rizzo and Whitman, 2009a).

As these considerations suggest, establishing that some bias observed in real-life situations should be corrected typically requires choice architects to determine what biases affect the targeted agents, these biases' impact on agents' behaviour, and the (cognitive, economic, motivational) costs involved in correcting such biases. Unfortunately, most studies control for the impact of only one or a few biases and limitations at a time (see e.g. Kagan, 2012, ch.1). Therefore, those studies' results rarely enable the new paternalists to anticipate the welfare implications of paternalistic interference in situations where a number of biases and limitations influence agents' behaviour simultaneously. To put it differently, more detailed evidence about the aetiology and the interactions of individuals' biases and limitations is needed to bridge the gap between the results obtained in controlled experimental settings and the new paternalists' claims concerning the welfare implications of their interventions.

(3) Suppose that the new paternalists could obtain accurate information concerning the short-term impact that specific biases and limitations have on agents' behaviour. Assume further that the new paternalists were able to calibrate their interventions for the interactions between these biases and limitations. Even this may not enable the new paternalists to design and implement welfare-enhancing paternalistic interventions. Doing so, in fact, would often require the new paternalists to anticipate *how responsive* agents' biases and limitations will be to paternalistic interference. Unfortunately, the responsiveness of several biases and limitations varies across time, agents and types of intervention in ways that are hard to quantify accurately (see e.g. Weinstein and Klein, 2002, on the resistance of personal risk perceptions to debiasing measures). Moreover, various paternalistic interventions have only short-term and context-dependent effects on agents' behaviour (see e.g. Stijn et al., 2010, on the impact of various traveller advisory systems). These complications do not prevent the new paternalists from providing approximate estimates of the welfare implications of particular interventions. Nonetheless, they considerably constrain the generalizability of the short-term results of paternalistic interventions targeting small population segments to longer time spans and wider subsets of the population (see e.g. Bonell et al., 2011, on paternalistic interventions in the health care system).

These generalizability concerns exacerbate when one considers that several types of interference championed by the new paternalists have been shown to backfire against their proponents' declared policy aims. To see this, take again 'save more tomorrow' plans, which aim to increase employees' savings for retirement by changing their default options in retirement saving decisions. The

hitherto implemented plans frequently tend to reduce (rather than increase) the targeted agents' retirement savings by inducing these agents to stick to default contributions that are lower than those agents' contributions under opt-in plans. For instance, as documented by Bubb and Pildes (2014), major 401(k) plans administrators have reported massive increases in the fraction of administered 401(k) plans and simultaneous decreases in both the median and the average total contribution rate of eligible employees. That is to say, pace leading new paternalists, too many and overly speculative inferential steps are required to accurately estimate the long-term responsiveness of agents' biases and limitations across agents and types of intervention.

(4) Individuals adopt several methods to alleviate the impact of specific biases and limitations on their own behaviour. Employed *self-regulative methods* range from self-imposed commitments to the voluntary submission to social controls and the advice of experts (see e.g. Trope and Fishbach, 2000). To be sure, individuals often lack proper incentives to de-bias and may fail to implement effective debiasing measures. Even so, self-regulative efforts can powerfully shape behaviour across several domains (see e.g. Baumeister and Vohs, 2004). For this reason, the new paternalists must accurately estimate the effects of individuals' self-regulative efforts in calibrating their interventions. Regrettably, several factors hamper this calibration task. To give one example, individuals differ in their propensity to self-regulation (see e.g. Carver and Scheier, 1998) and adopt dissimilar self-regulative methods whose efficacy varies across time and situations (see e.g. Bogg and Roberts, 2004). Moreover, the new paternalists frequently lack the means to ascertain to what extent observed behaviour is shaped by agents' self-regulative efforts. This, combined with the substitutability effects holding both between self-regulation at different times (see e.g. Baumeister et al., 1988) and between self-regulation and externally imposed controls (see e.g. Fishbach and Trope, 2005), often renders the estimation of the welfare implications of paternalistic interference prohibitively complicated (see e.g. the illustration regarding optimal sin taxes in point *1* above).

In light of all these concerns, a new paternalist may concede that the NCP faces significant epistemic and evidential challenges. At the same time, she may rebut that principled opposition to paternalism is "a literal nonstarter" on the alleged ground that choice architects cannot avoid providing defaults to the targeted agents and that there are *no viable alternatives* to paternalistic interference (Sunstein and Thaler, 2003, 1165; see also Sunstein, 2013a). This rebuttal does not insulate the NCP from the aforementioned epistemic and evidential challenges. For clearly, it is one thing to contend that choice architects typically make decisions (e.g. what information to provide and how to frame it) that influence agents' behaviour. It is quite another thing to allege that any such influence is bound to be paternalistic. To put it differently, the new paternalists' claims regarding the purported unavoidability of paternalistic interference seem to presuppose an implausibly broad conceptualization of paternalism.

A proponent of the NCP may acknowledge that the new paternalists often lack the evidence to establish what welfare implications paternalistic interference has for any *particular* agent. At the same time, she may object that the new paternalists can estimate the welfare implications of paternalistic interference

for distinct *types* of agents on the basis of assumptions concerning this interference's impact on the behaviour of such agents (see e.g. Sunstein, 2015). Suppose, for the sake of argument, that the new paternalists can neatly separate distinct types of agents (e.g. risk prone and risk averse agents) in terms of the behavioural impact some interference has on these agents. Assume further that the new paternalists can identify precisely which of their target agents belong to each behavioural type. Substantiating the NCP would require the new paternalists to establish systematic correspondences between the predicted *behavioural* impact and the putative *welfare* implications of interference for the targeted agents. Unfortunately, the alleged fact that the new paternalists are able to categorize their target agents into distinct behavioural types falls short of implying that they can also establish what welfare implications their interference has for such agents. For a given interference may simultaneously have a similar impact on the behaviour of some set of agents, and yet have dissimilar welfare implications for each of these agents. In fact, the concerns explicated in points *1-4* above provide compelling reasons to think that the new paternalists' interventions usually have dissimilar welfare implications for distinct individuals of the same behavioural types.

A new paternalist may further attempt to defend the NCP by pointing to wider *distributive considerations*. One such defence (see e.g. Guala and Mittone, 2015, and Trout, 2009, ch.4-6) goes as follows. Paternalistic interventions are routinely designed to redress injustice (e.g. undeserved imbalances due to genetic inheritance) and reduce the welfare losses that some agents' activities impose on third parties (e.g. think of compulsory health insurance schemes aimed at contrasting individuals' moral hazard). This does not *per se* render these paternalistic interventions justified. However, it forces anti-paternalists to specify what violations of agents' autonomy and consent would have to be present to license the claim that those interventions are unjustifiable. Now, distributive considerations may occasionally bear in favour of paternalistic interference. Nonetheless, there are at least three reasons to doubt that these considerations yield significant support to the NCP. First, there is no principled reason to expect that paternalistic interventions prevalently redress (rather than exacerbate) injustice and reduce (rather than increase) the welfare losses that their target agents' activities impose on third parties. Second, choice architects frequently have strong incentives to make decisions that promote their own interests (or the interests of third parties such as private firms) at the expense of their target agents (see e.g. Klick and Mitchell, 2006, and Willis, 2013, for illustrations). And third, choice architects can often design non-paternalistic interventions that redress injustice and reduce welfare losses without violating agents' autonomy or consent (see e.g. Anderson, 2010, on various attempts to solve collective action problems by engaging individuals in interpersonal deliberation rather than subjecting them to paternalistic interference).

At this stage, a new paternalist may rebut that the merits of the NCP should be judged on a *case-by-case* basis (see e.g. Sunstein, 2013b), and that choice architects should implement only those paternalistic interventions that meet the proffered justificatory challenges (see e.g. Sunstein, 2015, for a reply to some autonomy-related challenges). This rebuttal qualifies former calls for paternalistic interference, but provides rather limited support to the NCP. For *in primis*, confining the NCP's applicability to specific cases constitutes a

significant downplaying of the new paternalists' original ambition to systematically influence policy agendas across a variety of domains (see e.g. Thaler and Sunstein, 2008, ch.16-19). And second, the justificatory challenges articulated in this paper target not just a few selected instances of paternalistic interference, but also paradigmatic types of interference championed by many new paternalists. In this respect, it would be of little import to appeal to *collect more findings* about the aetiology and the impact of specific decision biases to identify additional ways to enhance agents' well-being (see e.g. Trout, 2005). Indeed, these appeals could even backfire against the new paternalists. For such findings may provide choice architects with more effective means to improve agents' well-being without having to implement any paternalistic interference, and may enable agents themselves to adopt superior forms of self-regulation, thereby reducing the putative need for any welfare-enhancing interference (see e.g. Gigerenzer, 2015, on how improving individuals' statistical skills enables them to contrast framing manipulations; see also Voorhoeve, 2013, on how behavioural and psychological findings increase agents' ability to enhance their own well-being through pre-commitment strategies).

Conclusion

The new paternalists maintain that the evidence collected in distinct decision sciences enables them to design and implement paternalistic interventions that address traditional anti-paternalistic objections and reliably enhance the well-being of their target agents. This new case for paternalism supplements previous calls in favour of paternalism with a wide array of behavioural and neuro-psychological findings. However, it justifies a much narrower range of paternalistic interventions than the new paternalists allege. Furthermore, the new paternalists' attempts to show that the paternalistic interventions they advocate are justified face severe and hitherto unaddressed justificatory challenges. In this article, I articulated and defended three such challenges in turn. More specifically, the *argument from conceptual ambiguity* documents the new paternalists' need to provide more precise and plausible criteria for evaluating the welfare implications of their interventions. The *argument from limited overlap* challenges the new paternalists to demonstrate that the interventions they advocate reliably enhance the well-being of their target agents without involving morally objectionable violations of these agents' autonomy or consent. The *argument from constrained epistemic access* illustrates that the new paternalists typically lack the information required to design and implement welfare-enhancing paternalistic interventions. These three arguments do not license all-encompassing opposition to paternalism. Still, taken together, they cast serious doubts on the new paternalists' claim that recent findings from the decision sciences provide convincing reasons for paternalistic interference.

REFERENCES

Anderson, J. 2010. Review of Richard Thaler and Cass Sunstein: nudge: improving decisions about health, wealth, and happiness. *Economics and Philosophy*, 26, 369–376.

Archard, D. 1993. Self-justifying paternalism. *Journal of Value Inquiry*, 27, 341-352.

Arneson, R.J. 1980. Mill Versus Paternalism. *Ethics*, 90, 470-489.

Arneson, R.J. 1999. Human Flourishing versus Desire Satisfaction. *Social Philosophy and Policy*, 16 (1), 113-142.

Arneson, R.J. 2005. Joel Feinberg and the Justification of Hard Paternalism. *Legal Theory*, 11 (3), 259-284.

Barber, B.M. and Odean, T. 2001. Boys Will Be: Gender, Overconfidence, and Common Stock Investment. *The Quarterly Journal of Economics*, 116, 261-292.

Bar-Gill, O. and Sunstein, C.R. 2015. Regulation as Delegation. *Journal of Legal Analysis*, 7 (1), 1-36.

Baumeister, R. 2002. Yielding to Temptation: Self-Control Failure, Impulsive Purchasing, and Consumer Behavior. *The Journal of Consumer Research*, 28 (4), 670-676.

Baumeister, R., Bratslavsky, E., Muraven, M. and Tice, D. 1998. Ego Depletion: is the Active Self a Limited Resource? *Journal of Personality and Social Psychology*, 74 (5), 1252-1265.

Baumeister, R. and Vohs, K.D. 2004. *Handbook of self-regulation: research, theory, and applications.* Guilford.

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U. and Fehr, E. 2008. Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron*, 58 (4), 639-650.

Bernheim, B.D. 2009. On the potential of neuroeconomics: A critical (but hopeful) appraisal. *American Economic Journal: Microeconomics*, 1, 1-41.

Bernheim, B.D. and Rangel, A. 2007. Toward choice-theoretic foundations for behavioral welfare economics. *American Economic Review*, 97, 464-470.

Bhargava, S. and Loewenstein, G. 2015. Behavioral Economics and Public Policy 102: Beyond Nudging. *American Economic Review*, 105, 396-401.

Bogg, T. and Roberts, B.W. 2004. Conscientiousness and Health-Related Behaviors: a Meta-Analysis of the Leading Behavioral Contributors to Mortality. *Psychological Bulletin*, 130, 887-919.

Bonell, C., McKee, M., Fletcher, A., Wilkinson, P. and Haines, A. 2011. One Nudge Forward, Two Steps Back. *British Medical Journal*, 342, d401.

Bovens, L. 2009. The Ethics of *Nudge.* In Grüne-Yanoff, T. and Hansson, S.O. *Preference Change: Approaches from Philosophy, Economics and Psychology*, Springer, Ch.10.

Bubb, R. and Pildes, R. 2014. How Behavioral Economics Trims Its Sails and Why. *Harvard Law Review*, 127, 1593–1678.

Bullock, E.C. 2015. A normatively neutral definition of paternalism. *Philosophical Quarterly*, 65 (258), 1-21.

Camerer, C.F. 2006. Wanting, Liking, and Learning: Neuroscience and Paternalism. *University of Chicago Law Review*, 73 (1), 87-110.

Camerer, C.F., Issacharoff, S., Loewenstein, G., O'Donoghue, T. and Rabin, M. 2003. Regulation for conservatives: behavioral economics and the case for asymmetric paternalism. *University of Pennsylvania Law Review*, 151, 1211-1254.

Carter, I. 1999. *A Measure of Freedom*. Oxford University Press.

Carter, I. 2014. Is the capability approach paternalist? *Economics and Philosophy*, 30, 75–98.

Carver, C.S. and Scheier, M.F. 1998. *On the self-regulation of behavior.* Cambridge University Press.

Cohen, S. 2013. Nudging and Informed Consent. *The American Journal of Bioethics*, 13 (6), 3-11.

Crisp, R. 2006. Hedonism Reconsidered. *Philosophy and Phenomenological Research*, 73, 619-645.

Darwall, S. 2006. The Value of Autonomy and the Autonomy of the Will. *Ethics*, 116, 263–84.

De Marneffe, P. 2006. Avoiding Paternalism. *Philosophy and Public Affairs*, 34 (1), 68-94.

Diener, E. and Seligman, M. 2004. Beyond money. Toward an economy of well-being. *Psychological Science in the Public Interest*, 5, 1–31.

Dworkin, G. 1972. Paternalism. *Monist*, 56, 64-84.

Dworkin, G. 1983. Paternalism: Some Second Thoughts. In *Paternalism*, Sartorius, R. (Ed.). University of Minnesota Press, 19-34.

Dworkin, G. 1988. *The Theory and Practice of Autonomy*. Cambridge University Press.

Dworkin, G. 2005. Moral Paternalism. *Law and Philosophy*, 24, 305-319.

Dworkin, G. 2010. Paternalism. *The Stanford Encyclopedia of Philosophy*, Zalta E.N. (Ed.). Available at: http://plato.stanford.edu/archives/sum2010/entries/paternalism/.

Elster, J. 1984. *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge University Press.

Farah, M.J. 2012. Neuroethics: the ethical, legal, and societal impact of neuroscience. *Annual Review Psychology*, 63, 571-91.

Feinberg, J. 1971. Legal paternalism. *Canadian Journal of Philosophy*, 1, 106-124.

Feinberg, J. 1986. *Harm to Self*. Oxford University Press.

Feldman, F. 1997. On the Intrinsic Value of Pleasures. *Ethics*, 107, 448-466.

Felsen, G., Castelo, N. and Reiner, P. 2013. Decisional enhancement and autonomy: public attitudes towards overt and covert nudges. *Judgment and Decision Making*, 8, 202-213.

Fishbach, A. and Trope, Y. 2005. The Substitutability of External Control and Self-Control. *Journal of Experimental Social Psychology*, 41, 256-270.

Gigerenzer, G. 2015. On the Supposed Evidence for Libertarian Paternalism. *Review of Philosophy and Psychology*, 6, 361-383.

Glaeser, E. 2006. Paternalism and Psychology. *University of Chicago Law Review*, 73, 133-156.

Griffin, J. 1986. *Well-Being: its Meaning, Measurement, and Moral Importance*. Oxford University Press.

Groll, D. 2012. Paternalism, Respect, and the Will. *Ethics*, 122 (4), 692-720.

Guala, F. 2005. *The methodology of experimental economics*. Cambridge University Press.

Guala, F. and Mittone, L. 2015. A political justification of nudging. *Review of Philosophy and Psychology*, 6 (3), 385-395.

Hausman, D.M. 2010. Hedonism and welfare economics. *Economics and Philosophy*, 26, 321-344.

Hausman, D.M. 2011. Mistakes about preferences in the social sciences. *Philosophy of the Social Sciences*, 41, 3–25.

Hausman, D.M. 2012. *Preference, value, choice, and welfare*. Cambridge University Press.

Hausman, D.M. 2016. On the Econ Within. *Journal of Economic Methodology*, 23, 26-32.

Hausman, D.M. and McPherson, M.S. 2009. Preference satisfaction and welfare economics. *Economics and Philosophy*, 25, 1-25.

Hausman, D.M. and Welch, B. 2010. Debate: To Nudge or Not to Nudge**.** *The Journal of Political Philosophy*, 18 (1), 123-136.

Husak, D. 2010. Paternalism and Consent. In Miller, F. and Wertheimer, A. (Eds.). *The Ethics of Consent. Theory and Practice*. Oxford University Press, 107-130.

Infante, G., Lecouteux, G. and Sugden, R. 2016. Preference purification and the inner rational agent: a critique of the conventional wisdom of behavioural welfare economics. *Journal of Economic Methodology*, 23, 1-25.

Jolls, C. and Sunstein, C.R. 2006. Debiasing Through Law. *Journal of Legal Studies*, 35 (1), 199-241.

Kagan, J. 2012. *Psychology's Ghosts*. Yale University Press.

Kagan, S. 1992. The limits of well-being. *Social Philosophy and Policy*, 9, 169–189.

Kant, I. 1797 [1996]. *MM. The Metaphysic of Morals*. Prussian Academy Volume VI. Transl. Gregor, M. Cambridge.

Klick, J. and Mitchell, G. 2006. Government Regulation of Irrationality: Moral and Cognitive Hazards. *Minnesota Law Review*, 90, 1620-1663.

Kosfeld, M., Heinrichs, M., Zak, P.J., Fischbacher, U. and Fehr, E. 2005. Oxytocin Increases Trust in Humans. *Nature,* 435, 673-676.

Le Grand, J. and New, B. 2015. *Government Paternalism: Nanny State or Helpful Friend?* Princeton University Press.

Loewenstein, G., Bryce, C., Hagmann, D. and Rajpal, S. 2015. Warning: You are about to be nudged. *Behavioral Science and Policy*, 1, 35-42.

Loewenstein, G. and Haisley, E. 2008. The Economist as Therapist: Methodological Ramifications of 'Light' Paternalism. In A. Caplin and Schotter, A. (Eds.). *Perspectives on the Future of Economics: Positive and Normative Foundations*.

Mill, J.S. 1859 [1956]. *On Liberty*. Bobbs-Merrill.

Mitchell, G. 2005. Libertarian Paternalism is an Oxymoron. *Northwestern University Law Review*, 99 (3), 1245-1277.

New, B. 1999. Paternalism and Public Policy. *Economics and Philosophy,* 15, 63–83.

Nussbaum, M.C. and Sen, A. 1993. *The Quality of Life*. Clarendon Press.

O'Donoghue, T. and Rabin, M. 2006. Optimal Sin Taxes. *Journal of Public Economics*, 90, 1825-1849.

Parfit, D. 1984. *Reasons and Persons*. Oxford Paperbacks.

Qizilbash, M. 2012. Informed desire and the ambitions of libertarian paternalism. *Social Choice and Welfare*, 38, 647–658.

Rachlinski, J. 2003. The Uncertain Psychological Case for Paternalism. *Northwestern University Law Review*, 97 (3), 1165-1225.

Rizzo, M.J. and Whitman, D.G. 2009a. Little Brother is Watching You: New Paternalism on the Slippery Slopes. *Arizona Law Review*, 51, 685-739.

Rizzo, M.J. and Whitman, D.G. 2009b. The Knowledge Problem of the New Paternalism. *Brigham Young University Law Review*, 103-161.

Rubinstein, A. and Arad, A. 2015. The People's Perspective on Libertarian-Paternalistic Policies. At: http://www.tau.ac.il/~aradayal/LP.pdf.

Rubinstein, A. and Salant, Y. 2012. Eliciting welfare preferences from behavioral datasets. *Review of Economic Studies*, 79, 375–387.

Salant, Y. and Rubinstein, A. 2008. (A, f): Choice with frames. *Review of Economic Studies*, 75, 1287-1296.

Samuelson, W. and Zeckhauser, R. 1988. Status Quo Bias in Decision Making. *Journal of Risk and Uncertainty*, 1, 7-59.

Shiffrin, S.V. 2000. Paternalism, Unconscionability Doctrine, and Accommodation. *Philosophy and Public Affairs*, 29, 205–250.

Sobel, D. 1998. Well-being as the Object of Moral Consideration. *Economics and Philosophy*, 14 (2), 249-281.

Stijn, D., Vanrie, J., Dreesen, A. and Brijs, T. 2010. Additional road markings as an indication of speed limits: Results of a field experiment and a driving simulator study. *Accident Analysis & Prevention*, 42 (3), 953-960.

Sugden, R. 2004. The opportunity criterion: consumer sovereignty without the assumption of coherent preferences. *American Economic Review,* 94, 1014-1033.

Sugden, R. 2006. Taking unconsidered preferences seriously. In *Preferences and well-being*. Olsaretti, S. (Eds.) Paperback, 209-232.

Sugden, R. 2007. The value of opportunities over time when preferences are unstable. *Social Choice and Welfare*, 29, 665-682.

Sugden, R. 2008. Why incoherent preferences do not justify paternalism. *Constitutional Political Economy*, 19, 226-248.

Sugden, R. 2013. The behavioural economist and the social planner: to whom should behavioural welfare economics be addressed? *Inquiry*, 56, 519–538.

Sumner, L.W. 1995. The Subjectivity of Welfare. *Ethics*, 105, 764-790.

Sunstein, C.R. 2013a. Deciding by Default. *University of Pennsylvania Law Review*, 162, 1-57.

Sunstein, C.R. 2013b. The Storrs Lectures: Behavioral Economics and Paternalism. *Yale Law Journal*. 122, 1826–1899.

Sunstein, C.R. 2015. Nudging and Choice Architecture: Ethical Considerations. *Yale Journal on Regulation*. At: http://www.law.harvard.edu/programs/olin_center/papers/pdf/Sunstein_809.pdf.

Sunstein, C.R. and Thaler, R.H. 2003. Libertarian paternalism is not an oxymoron. *University of Chicago Law Review*, 70 (4), 1159-1202.

Thaler, R.H. and Sunstein, C.R. 2003. Libertarian paternalism. *American Economic Review*, 93 (2), 175-179.

Thaler, R.H. and Sunstein, C.R. 2008. *Nudge: improving decisions about health, wealth, and happiness*. Yale University Press.

Trope, Y. and Fishbach, A. 2000. Counteractive self-control in overcoming temptation. *Journal of Personality and Social Psychology*, 79 (4), 493-506.

Trout, J.D. 2005. Paternalism and Cognitive Bias. *Law and Philosophy*, 24, 393-434.

Trout, J.D. 2009. *The Empathy Gap. Building Bridges to the Good Life and the Good Society*. Viking/Penguin.

Tversky, A. and Kahneman, D. 1974. Judgment Under Uncertainty: Heuristics and Biases. *Science*, 185, 1124-1131.

Velleman, J.D. 1999. A right to self-termination? *Ethics*, 109, 606-628.

Voorhoeve, A. 2013. Response to Rabin. In Oliver, A. (Ed.). *Behavioural Public Policy*. Cambridge University Press, 140-147.

Weinstein, N.D. and Klein, W.M. 2002. Resistance of Personal Risk Perceptions to Debiasing Interventions. In *Heuristics and Biases: The Psychology of Intuitive Judgmen*t. Gilovich, T., Griffin, D. and Kahneman, D. (Ed.). Cambridge University Press, 313-323.

White, M. 2013. *The Manipulation of Choice: Ethics and Libertarian Paternalism*. Palgrave Macmillan.

Wilkinson, T. M. 2013. Nudging and manipulation. *Political Studies*, 61 (2), 341–355.

Willis, L.E. 2013. When Nudges Fail: Slippery Defaults. *University of Chicago Law Review*, 80, 1115-1229.

Wilson, J. 2011. Why It's Time to Stop Worrying About Paternalism in Health Policy. *Public Health Ethics*, 4 (3), 269-279.

Zamir, E. 1998. The Efficiency of Paternalism. *Virginia Law Review*, 84 (2), 229-284.